Course Overview

Tal Linzen

CDS, NYU

Jan 25, 2022

Contents

Logistics

Course Staff

- Instructors:
  - Tal Linzen (Linguistics and Data Science)
  - He He (Computer Science and Data Science)
- Section leaders:
  - Vishakh Padmakumar
  - Colin Wan
- Graders:
  - Congyun Jin
  - Bella Lyu
  - Gavin Nan
  - Ziyi Xie
  - Namrata Mukhija

Logistics

- Class webpage: https://nyu-ds1003.github.io/spring2022
  - Course materials (lecture slides, homework assignments) will be made available on the website
- Announcements via Brightspace
- Discussion / questions on CampusWire: https://campuswire.com/c/G6A12AE75/feed
- Sign up to Gradescope to submit homework assignments (entry code **V8K3XW**)
- Office Hours:
  - The professors' office hours will be on Zoom (Tal: Tuesday 3:30-4:30 pm; He: Thursday 4:30-5:30 pm)
  - The section leaders' office hours will be in person (Vishakh: Wednesday 6-7 pm; Colin: Monday 1-2 pm; Room 204, 60 5th Ave)

Assessment

- 7 assignments ($1 \times 4\% + 6 \times 6\% = 40\%$)
- Two tests (60%)
  - Midterm Exam (30%) in Week 7 (March 8th), covering material up to Week 6
  - Final Exam (30%), schedule hasn't been announced yet, covering all material
- Typical grade distribution: A (40%), A- (20%), B+ (20%), B (10%), B- (5%), <B- (5%)

Homework

- Assignment 0: Help you get farmiliar with the format (not submitted or graded)
- First assignment out now – due on **Feb 1**
- Submit through Gradescope as a **PDF document**
- Late policy: Assignments are accepted up to **48 hours** late (see more details on website)
- You can collaborate with other students on the homework assignments, but please:
  - Write up the solutions and code on your own;
  - And list the names of the students you discussed each problem with.

Exams (60%)

- Exam format TBD: either in-person or submitted through Gradescope, like the assignments
- We'll make this decision based on NYU policy at the time – stay tuned
- Before each exam, we will post exams from previous years

Prerequisites

- DS-GA 1001: Introduction to Data Science
- DS-GA 1002: Statistical and Mathematical Methods
- Math
  - Multivariate Calculus
  - Linear Algebra
  - Probability Theory
  - Statistics
  - [Preferred] Proof-based linear algebra or real analysis
- Python programming (numpy)

Course Overview and Goals

Syllabus (Tentative)

13 weeks of instruction + 1 week midterm exam

- 2 weeks: introduction to **statistical learning theory**, **optimization**
- 2–3 weeks: **Linear** methods for binary classification and regression (also **kernel methods)**
- 2 weeks: **Probabilistic models**, **Bayesian** methods
- 1 week: **Multiclass** classification and introduction to **structured prediction**
- 3–4 weeks: **Nonlinear** methods (**trees**, **ensemble** methods, and **neural networks**)
- 2 weeks: **Unsupervised** learning: **clustering** and **latent variable** models
- More detailed schedule on the course website (still subject to change)
- Certain applications and practical algorithms may be covered in the labs

The high level goals of the class

- Our focus will be on the fundamental building blocks of machine learning
- ML methods have a lot of names; our goal is for you to notice that

  fancy new method A "is just" familiar thing B + familiar thing C + tweak D

  - SVM "**is just**" ERM with hinge loss with ℓ_2 regularization
  - Pegasos "**is just**" SVM with SGD with a particular step size rule
  - Random forests "**are just**" bagging with trees, with a different approach to choosing splitting variables

The level of the class

- We will learn how to implement each ML algorithm **from scratch** using numpy alone, without any ML libraries.
- Once we have implemented an algorithm from scratch once, we will use the sklearn version.